Introduction to Bayesian inference

End of course data analysis project

—

The Center for Disease Control (CDC) reports the vaccination coverage of Varicella among young children. Varicella, commonly known as chicken- pox, is a highly contagious viral infection caused by the varicella-zoster virus (VZV). Vaccination against chickenpox has been highly effective in reducing the incidence and severity of the disease. In the United States, vaccination against varicella has been part of the routine childhood immunization sched- ule since the mid-1990s. Since the vaccine's introduction, there has been a dramatic decline in the number of chickenpox cases, hospitalizations, and deaths associated with the disease. The target for vaccination coverage of varicella (chickenpox) in the United States, as set by the Centers for Dis- ease Control and Prevention (CDC), is typically around 90% or higher for children. This high coverage rate is aimed at achieving herd immunity and preventing outbreaks of chickenpox within communities.

Project 1: Insurance

The next table summarizes, based on a survey, the number of children in the birth cohort 2014-2017 that had at least one dose of the Varicella vaccine. It gives the number of vaccinated children (Vaccinated) amongst the number of children in the survey (Sample Size). The information is provided for 5 regions of the US, and split according to insurance status (private insurance, uninsured or any Medicaid).

Geography	Insurance	Vaccinated	Sample Size
North Carolina	Any Medicaid	380	419
North Carolina	Private Insurance Only	632	673
North Carolina	Uninsured	28	34
Georgia	Any Medicaid	363	396
Georgia	Private Insurance Only	527	576
Georgia	Uninsured	36	50
Wisconsin	Any Medicaid	282	332
Wisconsin	Private Insurance Only	514	548
Wisconsin	Uninsured	16	34
Florida	Any Medicaid	446	490
Florida	Private Insurance Only	588	628
Florida	Uninsured	28	39
Mississippi	Private Insurance Only	400	441
Mississippi	Uninsured	27	32

Question 1

> Derive analytically the posterior of the vaccination coverage per ge- ography and insurance group. Use a conjugate prior that (1) reflects no knowledge on the vaccination coverage, and (2) reflects that vac- cination coverage is typically around 90% or higher. Give posterior summary measures of the vaccination coverage per geography and in- surance group. Is the choice of the prior impacting your results?

Theoretical considerations

The outcome *Vaccinated/Not Vaccinated* follows a Bernouilli distribution with parameter p:

V : Vaccination status $V \in \{0, 1\}$ $V \sim \mathcal{B}ern(p)$

It is known from theory that the sum of n *i.i.d* Bernoulli random variables follows a Binomial distribution. This will be used to model the sample outcome: the number of vaccinated people V_s in a random sample of size n:

$$V_s = \sum_{i}^{n} V_i \sim \mathcal{B}inom(n, \theta)$$

where θ is the parameter of interest - the vaccine coverage.

In the course, we saw that the Beta distribution is the conjugate prior for binomially distributed data:

Table 2: Beta-Binomial conjugate model

Distribution	Formula
Prior	$p(\theta) = \mathcal{B}eta(\alpha, \beta)$
Likelihood	$p(y \mid \theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$
Posterior	$p(\theta \mid y) = \mathcal{B}eta(\alpha + k, \beta + n - k)$

The summary measures for the Beta distribution are defined as follows:

Table 3: Beta distribution summary measures

Summary Measure	Formula
Mean	$\frac{\alpha}{\alpha+\beta}$
Median	See Note
Mode	$\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$

Note: The median of the Beta distribution does not have a simple closed form expression. It can be approximated numerically or using statistical software.

Choice of prior distributions

(1) No prior knowledge

In order to reflect no prior knowledge on the vaccine coverage, the weakly-informative prior Beta(1,1) will be used, which is equivalent to the uniform distribution over $[0, 1]$.

(2) Vaccine coverage >90%

For modeling prior knowledge that vaccine coverage is about 90%, we chose the Beta(150, 7) distribution.

Comparison of priors

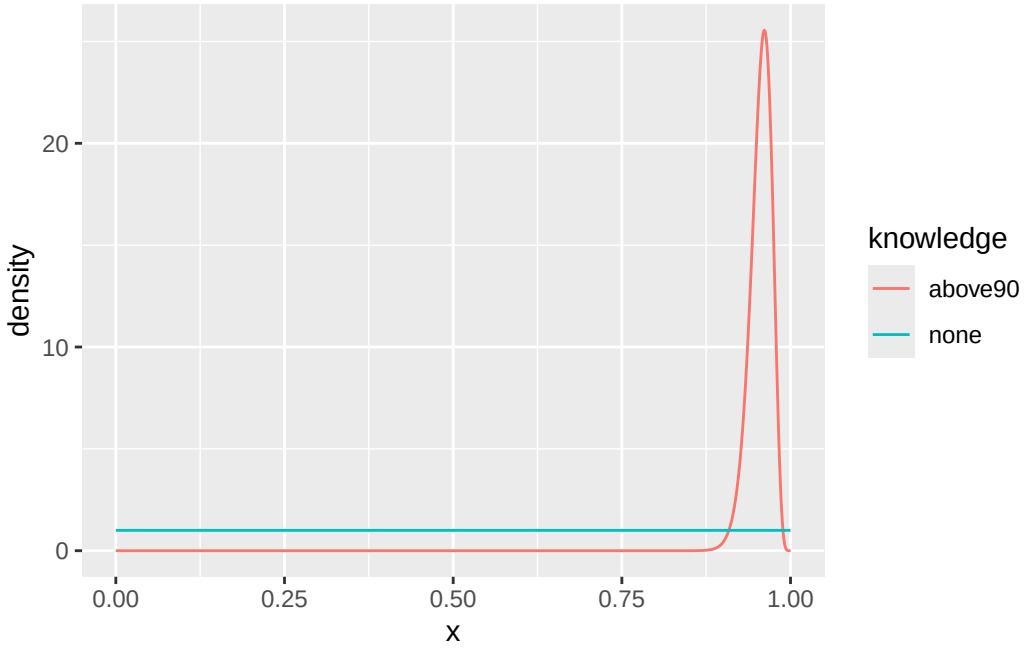


Figure 1: Comparison of uniform Beta(1,1) prior vs. strong Beta(150,7) prior

Results

Table 4: Posterior distribution parameters and summary measures per geography for the two different Beta priors

		Posterior parameters and summary measures												
		Beta(1,1) prior							Beta(150,7) prior					
Geography	Insurance	α	β	mean	mode	var	HPD LL	HPD UL	α	β	mean	mode	var	HPD LL	HPD UL
NC	Medicaid	381	40	0.905	0.907	0.014	0.875	0.931	530	46	0.920	0.922	0.011	0.897	0.941
NC	Private	633	42	0.938	0.939	0.009	0.918	0.955	782	48	0.942	0.943	0.008	0.925	0.957
NC	Uninsured	29	7	0.806	0.824	0.065	0.664	0.916	178	13	0.932	0.937	0.018	0.892	0.963
GA	Medicaid	364	34	0.915	0.917	0.014	0.885	0.940	513	40	0.928	0.929	0.011	0.905	0.948
GA	Private	528	50	0.913	0.915	0.012	0.889	0.935	677	56	0.924	0.925	0.010	0.903	0.942
GA	Uninsured	37	15	0.712	0.720	0.062	0.583	0.825	186	21	0.899	0.902	0.021	0.854	0.936
WI	Medicaid	283	51	0.847	0.849	0.020	0.807	0.884	432	57	0.883	0.885	0.014	0.854	0.910
WI	Private	515	35	0.936	0.938	0.010	0.915	0.955	664	41	0.942	0.943	0.009	0.923	0.958
WI	Uninsured	17	19	0.472	0.471	0.082	0.314	0.634	166	25	0.869	0.873	0.024	0.818	0.913
FL	Medicaid	447	45	0.909	0.910	0.013	0.882	0.932	596	51	0.921	0.922	0.011	0.899	0.941
FL	Private	589	41	0.935	0.936	0.010	0.914	0.953	738	47	0.940	0.941	0.008	0.923	0.956
FL	Uninsured	29	12	0.707	0.718	0.070	0.561	0.834	178	18	0.908	0.912	0.021	0.864	0.944
MS	Private	401	42	0.905	0.907	0.014	0.876	0.931	550	48	0.920	0.921	0.011	0.897	0.940
MS	Uninsured	28	6	0.824	0.844	0.064	0.681	0.930	177	12	0.937	0.941	0.018	0.898	0.967

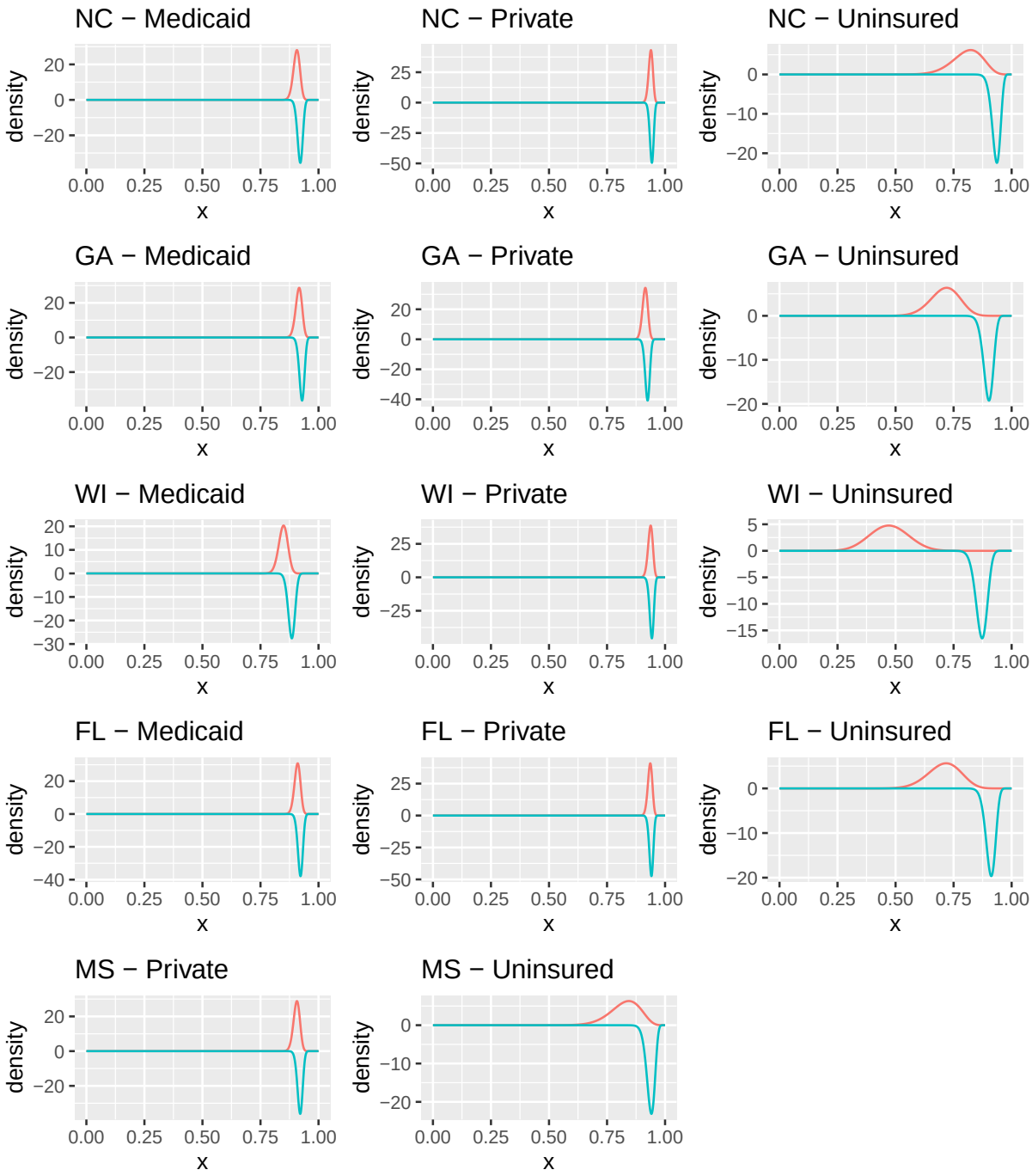


Figure 2: Posterior distributions per Geography/Insurance stratum, positive density: Beta(1,1) prior, negative density: Beta(150,7)

Question 2

> nvestigate whether the vaccination coverage is associated with the in- surance status using a logistic regression model Yij Binom(ij ,Nij) with logit(ij) = 0 + 1IAnyMedicaid,ij + 2IUninsured,ij where i is the location, j is the insurance status, ij is the vaccination coverage and I. are dummy variables. Assume non-informative priors for the parameters to be estimated. Write and explain the code in BUGS language

```
model
{
  for (t in 1:T) {
    # Likelihood
```

```
    y[t] ~ dbin(p[t], K[t])
    # conditional mean model using link function
    logit(p[t]) <- alpha_0 + ins_1 * x_1[t] + ins_2 * x_2[t]
 }

  # Priors
  alpha_0 ~ dnorm(0.0,0.01)
  ins_1 ~ dnorm(0.0,0.01)
  ins_2 ~ dnorm(0.0,0.01)


  # Vaccine coverage per Insurance group
  pi_private <- exp(alpha_0)/(1+exp(alpha_0))
  pi_medicaid <- exp(alpha_0 + ins_1)/(1+exp(alpha_0 + ins_1))
  pi_uninsured <- exp(alpha_0 + ins_2)/(1+exp(alpha_0 + ins_2))

  # Difference between vaccine coverage per insurance group
  diff_priv_medicaid  <- pi_private - pi_medicaid
  diff_priv_uninsured <- pi_private - pi_uninsured
}
```

Likelihood function and model specification

For each row `t` in the dataset, the outcome `y[t]` (number of vaccinated children in the sample) is specified as drawn from a Binomial distribution with parameters `p[t]` and sample size `K[t]`. We also specify the conditional mean model for this likelihood function using the `logit` link function.

Prior distribution

As prior distribution for all parameters (intercept and indicator variables for the insurance group), a vague prior is chosen : $\mathcal{N}(\mu = 0, \tau = 0.01)$.

Quantities of interest During each MCMC run, the vaccine coverage per insurance group is calculated using the inverse logit transformation, this will give access to the posterior distribution of vaccine coverage per insurance group. Furthermore, the differences between vaccine coverages are calculated which will be needed to answer Question 6.

Question 3

> Run the MCMC method and check convergence of the MCMC chains. Give the details on how you checked convergence.

MCMC run summary

```
Inference for Bugs model at "4", fit using jags,
 2 chains, each with 40000 iterations (first 2000 discarded), n.thin = 2
 n.sims = 38000 iterations saved. Running time = 1.829 secs
                   mu.vect sd.vect   2.5%     25%     50%     75%   97.5%
alpha_0              2.565   0.073  2.425   2.516   2.564   2.613   2.711
alpha_1             -0.381   0.110 -0.598  -0.455  -0.381  -0.308  -0.165
alpha_2             -1.642   0.178 -1.985  -1.763  -1.645  -1.524  -1.286
diff_priv_medicaid   0.030   0.009  0.013   0.024   0.030   0.036   0.048
diff_priv_uninsured  0.214   0.033  0.150   0.191   0.213   0.236   0.280
pi_medicaid          0.899   0.007  0.883   0.894   0.899   0.904   0.913
pi_private           0.928   0.005  0.919   0.925   0.928   0.932   0.938
pi_uninsured         0.714   0.033  0.649   0.692   0.715   0.737   0.777
deviance           102.277   2.474 99.480 100.464 101.623 103.403 108.703
                    Rhat n.eff
alpha_0            1.002  2400
alpha_1            1.002  1300
alpha_2            1.001  4800
```

```
diff_priv_medicaid   1.002  1300
diff_priv_uninsured 1.001 10000
pi_medicaid          1.002  2600
pi_private           1.002  2500
pi_uninsured         1.001 24000
deviance             1.001 22000

For each parameter, n.eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor (at convergence, Rhat=1).

DIC info (using the rule: pV = var(deviance)/2)
pV = 3.1 and DIC = 105.3
DIC is an estimate of expected predictive error (lower deviance is better).
```

Convergence checks

Traceplots

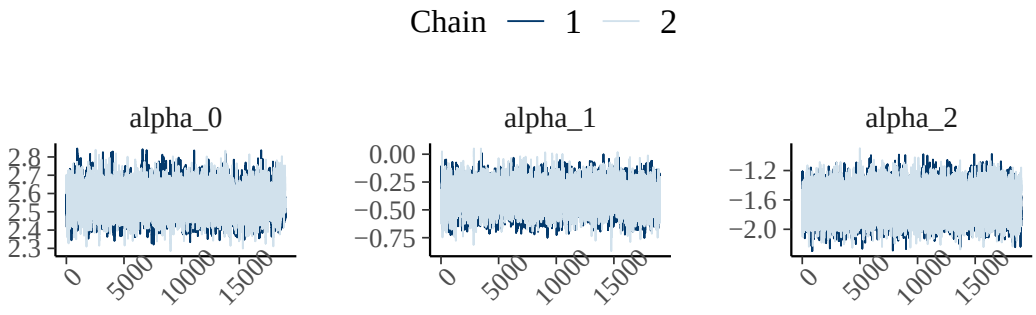


Figure 3

Density plots per chain

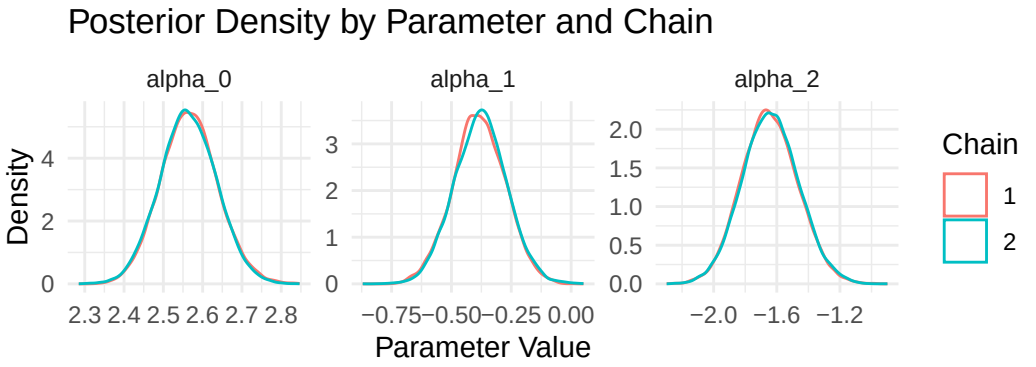


Figure 4

Autocorrelation plot

$\hat{R}$

```
Potential scale reduction factors:

          Point est. Upper C.I.
alpha_0            1          1
alpha_1            1          1
alpha_2            1          1

Multivariate psrf
```

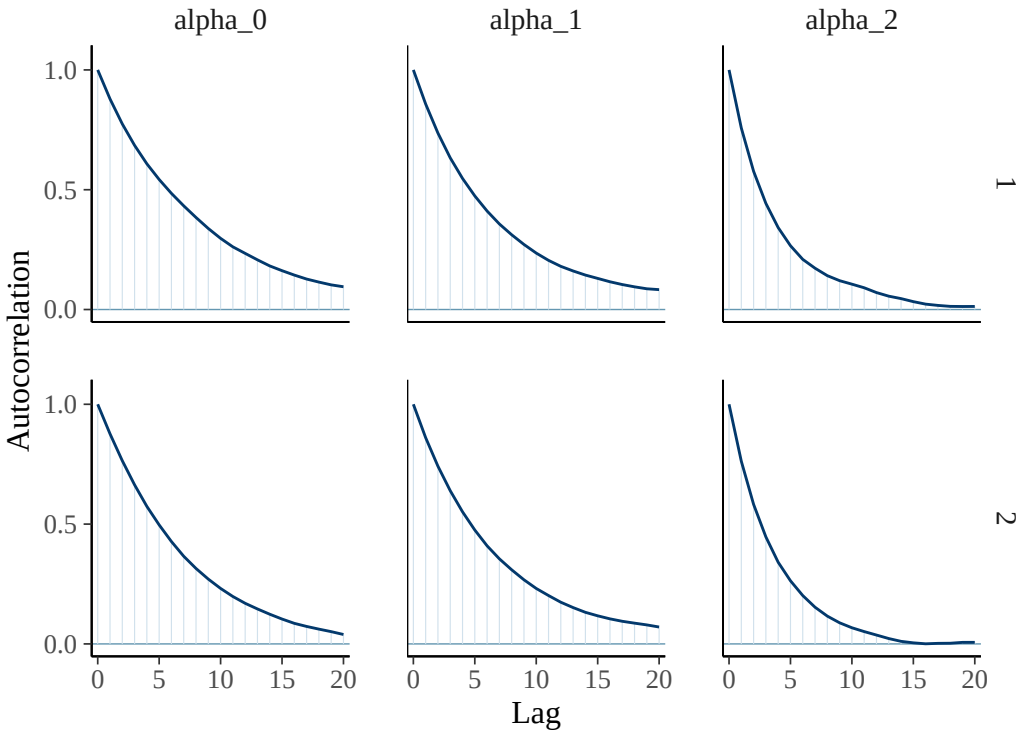


Figure 5

```
1
```

Geweke diagnostics

```
[[1]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

alpha_0 alpha_1 alpha_2
 0.3110 -0.1995 -0.0623


[[2]]

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

alpha_0 alpha_1 alpha_2
-0.5910  1.0079  0.5327
```

Question 4

> Make a plot of the posterior of the model parameters and give posterior summary measures. Interpret the results.

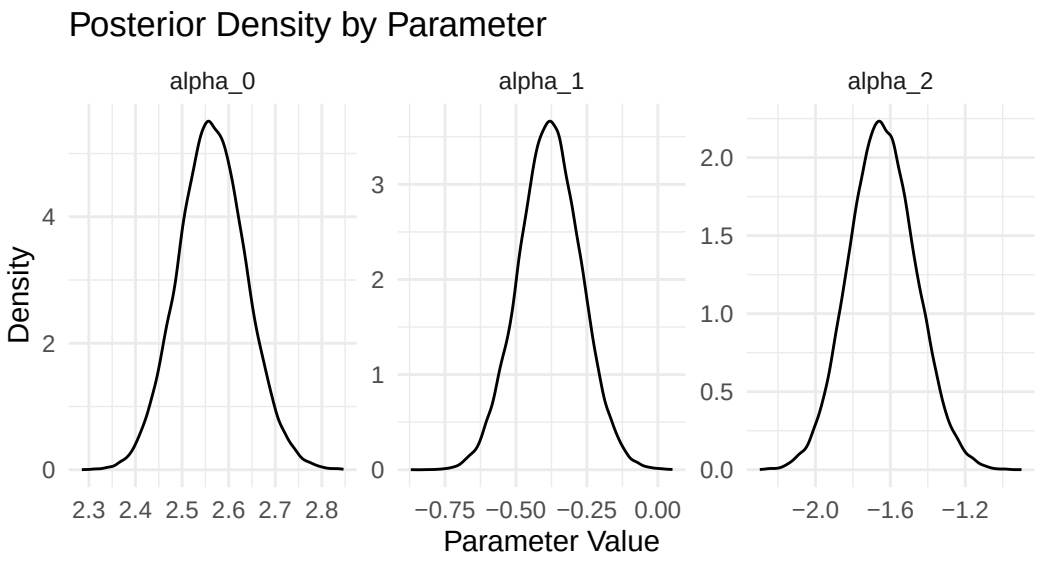


Figure 6

Table 5

Parameter	Summary measures			95% HPD interval	
	Mean	Median	SD	LL	UL
α_0	2.565	2.564	0.073	2.419	2.704
α_1	-0.381	-0.381	0.110	-0.598	-0.166
α_2	-1.642	-1.645	0.178	-1.991	-1.294

Plots of posterior densities and summary measures of the model parameters are given in Figure 6 and Table 5, respectively.

With `Private Insurance Only` as reference category the interpretation of the parameters is as follows:

- α_0 gives the $log(odds)$ of Vaccinated *vs.* Non-Vaccinated in the private insurance group
- α_1 gives the change in $log(odds)$ in the Medicaid group *vs.* the private Insurance group, the $log(odds)$ in the Medicaid group are given by $\alpha_0 + \alpha_1$
- α_2 gives the change in $log(odds)$ in the Uninsured group *vs.* the private Insurance group, the $log(odds)$ in the Uninsured group are given by $\alpha_0 + \alpha_2$

Numerically, the posterior mean of α_0 is estimated at 2.565 and with a posterior probability of 95% α_0 lies in $[2.419, 2.704]$. The posterior estimates for α_1 and α_2 can be read in the same manner from Table 5.

Question 5

> Give the posterior estimate of the vaccination coverage per region and insurance status. Compare with the analytical results you obtained in Question 1.

The vaccine coverage in a given group can be obtained from the $log(odds)$ (see model code Question 3):

$$\pi = \frac{\exp(log(odds))}{1 + \exp(log(odds))}$$

Posterior estimates for the mean vaccine coverage per region and insurance status are given in 6. To facilite comparison against the results of conjugate modeling in Question 1, Figure 7 shows the difference of the posterior mean obtained from the logistic regression model ($\bar{\pi}_{logreg}$) with respect to a $Beta(1,1)$ prior ($\Delta_{Beta(1,1)}$), and to a $Beta(150,7)$ ($\Delta_{Beta(150,7)}$), respectively.

For both priors, the differ between in the posterior mean between the logistic regression model and the conjugate modeling is most pronounced for the group of Uninsured children. Compared to the strong prior, $\Delta_{Beta(150,7)}$ is negative in the Uninsured group across all states, while for the non-informative prior, $\Delta_{Beta(1,1)}$ is negative for uninsured children in MS and NC, strongly positive in WI, and neglegible in FL and GA.

Table 6

Table 7: Posterior estimates of mean vaccine coverages from the logistic regression model vs. estimates from conjugate pair modeling. Subscripts indicate whether the estimates was obtained from logistic regression or the chosen prior distribution, respectively. $\Delta_{Beta(\alpha,\beta)}$ gives the difference between posterior estimates from the logistic regression model and conjugate pair modeling. Vaccine coverage for MS/Medicaid can only be obtained from the logistic regression model.

Geo.	Ins.	k	n	$\bar{\pi}_{logreg}$	$\bar{\pi}_{Beta(1,1)}$	$\bar{\pi}_{Beta(150,7)}$	$\Delta_{Beta(1,1)}$	$\Delta_{Beta(150,7)}$
FL	Medicaid	446	490	0.899	0.909	0.921	-0.010	-0.022
FL	Private	588	628	0.928	0.935	0.940	-0.007	-0.012
FL	Uninsured	28	39	0.714	0.707	0.908	0.007	-0.194
GA	Medicaid	363	396	0.899	0.915	0.928	-0.016	-0.029
GA	Private	527	576	0.928	0.913	0.924	0.015	0.004
GA	Uninsured	36	50	0.714	0.712	0.899	0.002	-0.185
MS	Medicaid	NA	NA	0.899	NA	NA	NA	NA
MS	Private	400	441	0.928	0.905	0.920	0.023	0.008
MS	Uninsured	27	32	0.714	0.824	0.937	-0.110	-0.223
NC	Medicaid	380	419	0.899	0.905	0.920	-0.006	-0.021
NC	Private	632	673	0.928	0.938	0.942	-0.010	-0.014
NC	Uninsured	28	34	0.714	0.806	0.932	-0.092	-0.218
WI	Medicaid	282	332	0.899	0.847	0.883	0.052	0.016
WI	Private	514	548	0.928	0.936	0.942	-0.008	-0.014
WI	Uninsured	16	34	0.714	0.472	0.869	0.242	-0.155

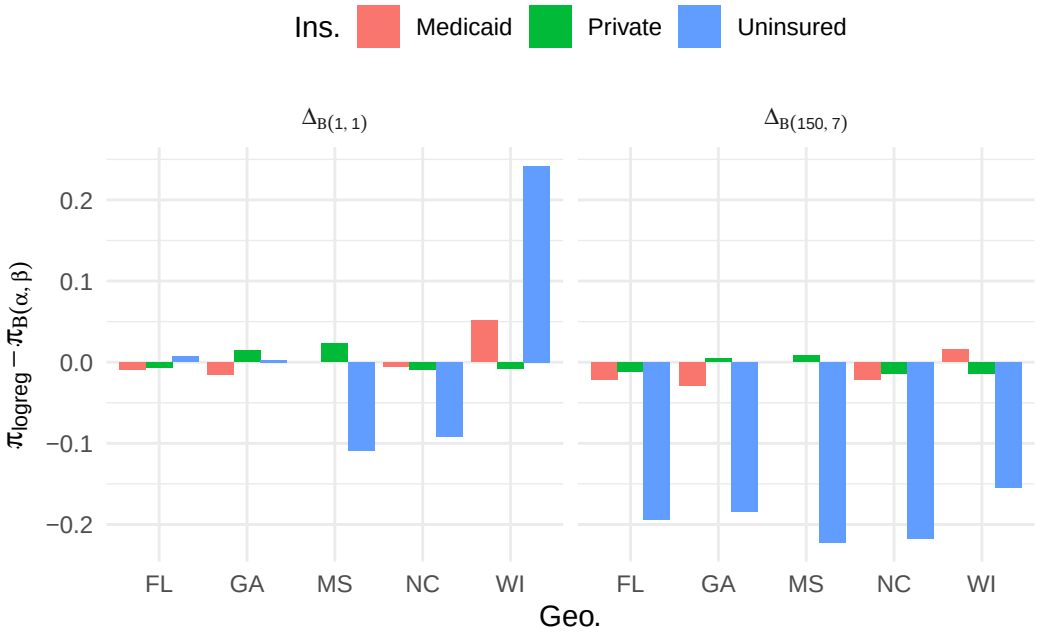


Figure 7: Differences between posterior means of vaccine coverage per region and insurance status obtained from logistic regression and conjugate pair modeling. Differences are most pronounced for the Uninsured group in MS, NC, WC when using a $Beta(1,1)$ prior and all states when using a $Beta(150,7)$ prior.

Question 6

> Based on the logistic regression model, what is the probability (a posteriori) that coverage amongst children that have private insurance is higher than amongst children that have any medicaid? And compared to children with no insurance?

To answer this question, the differences $\pi_{private} - \pi_{medicaid}$ and $\pi_{private} - \pi_{uninsured}$ were incorporated and observed during the MCMC run of the model specified in Question 3. Posterior densities and empirical CDF are shown in Figure 8.

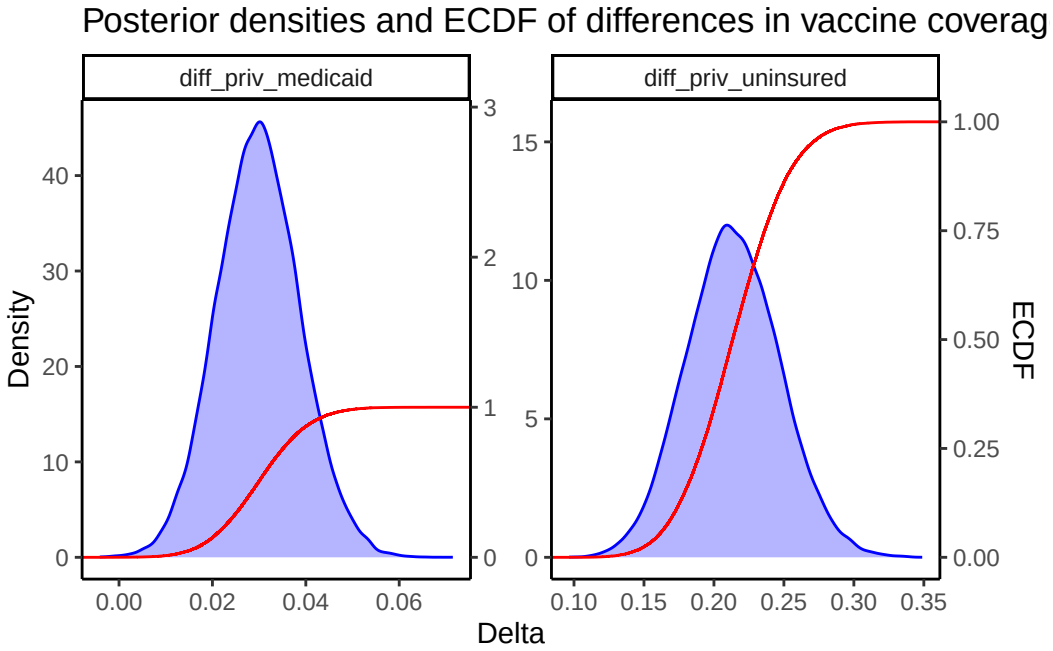


Figure 8

The probabilities of interest defined below and can be approximated by using the posterior samples from the MCMC runs :

1. $P(\pi_{priv} > \pi_{medicaid}) = P(\pi_{priv} - \pi_{medicaid} > 0) = 0.9995$
2. $P(\pi_{priv} > \pi_{uninsured}) = P(\pi_{priv} - \pi_{uninsured} > 0) = 1$

Question 7

> Secondly, investigate whether the vaccination coverages are distinct at the different locations by adding a location-specific intercept.
>
> $$\text{logit}\left(\pi_{ij}\right) = \alpha_{0i} + \alpha_1 I_{\text{AnyMedicaid}} + \alpha_2 I_{\text{Uninsured}}$$
>
> Assume non-informative priors for the parameters to be estimated. Write the code in BUGS language. Give a brief summary of the convergence checks you performed. Compare posteriors of vaccination coverages with results from Question 1.

```
model
{
  # Likelihood
  for (t in 1:T) {
    y[t] ~ dbin(p[t], K[t])
    logit(p[t]) <- inprod(beta, X[t,])
 }
```

```
  # Priors
  for (i in 1:7){
    beta[i] ~ dnorm(0.0,0.01)
  }

  # Vaccine coverage
  for (t in 1:T){
    #logodds
    lo[t] <- inprod(beta, X[t,])
    #convert to probability / vaccine coverage
    coverage[t] <- exp(lo[t]) / (1 + exp(lo[t]))
  }
}
```

Convergence was checked in a similar way as for the model defined in Question 3. In brief, trace plots showed the expected caterpillar pattern, posterior densities per parameter and chain showed superposed well and autocorrelation plots revealed decreasing autocorrelation with increasing lag number - all indicating convergence.

Table 8

Table 9: Posterior estimates of mean vaccine coverages from the logistic regression model including a region specific intercept vs. estimates from conjugate pair modeling. Subscripts indicate whether the esimates was obtained from logistic regression or the chosen prior distribution, respectively. $\Delta_{Beta(\alpha,\beta)}$ gives the difference between posterior estimates from the logistic regression model and conjugate pair modeling. Vaccine coverage for MS/Medicaid can only be obtained from the logistic regression model.

Geo.	Ins.	k	n	$\bar{\pi}_{logreg}$	$\bar{\pi}_{Beta(1,1)}$	$\bar{\pi}_{Beta(150,7)}$	$\Delta_{Beta(1,1)}$	$\Delta_{Beta(150,7)}$
FL	Medicaid	446	490	0.907	0.909	0.921	-0.001	-0.014
FL	Private	588	628	0.937	0.935	0.940	0.002	-0.003
FL	Uninsured	28	39	0.742	0.707	0.908	0.035	-0.166
GA	Medicaid	363	396	0.896	0.915	0.928	-0.018	-0.032
GA	Private	527	576	0.929	0.913	0.924	0.016	0.005
GA	Uninsured	36	50	0.717	0.712	0.899	0.005	-0.182
MS	Medicaid	NA	NA	0.881	NA	NA	NA	NA
MS	Private	400	441	0.918	0.905	0.920	0.013	-0.001
MS	Uninsured	27	32	0.686	0.824	0.937	-0.138	-0.251
NC	Medicaid	380	419	0.911	0.905	0.920	0.006	-0.009
NC	Private	632	673	0.940	0.938	0.942	0.002	-0.002
NC	Uninsured	28	34	0.751	0.806	0.932	-0.055	-0.181
WI	Medicaid	282	332	0.872	0.847	0.883	0.025	-0.011
WI	Private	514	548	0.912	0.936	0.942	-0.024	-0.030
WI	Uninsured	16	34	0.667	0.472	0.869	0.195	-0.202

Question 8

Compare the vaccination coverage in each of the location with the vaccination coverage in North Carolina:

$$\theta_{ij} = \frac{\pi_{ij}}{\pi_{\text{North Carolina },j}}$$

Interpret the results.

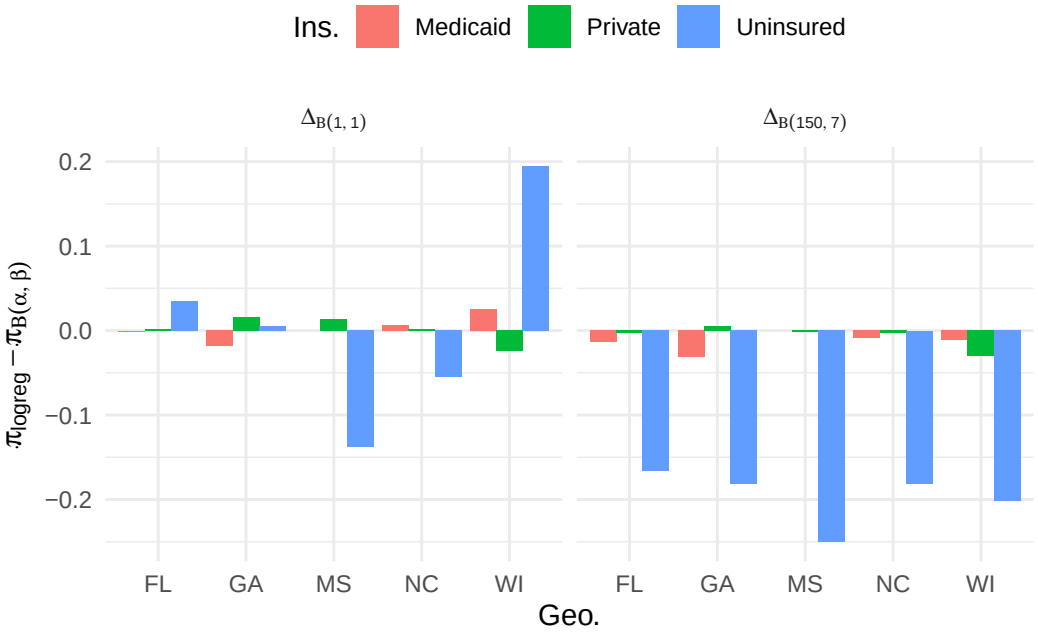


Figure 9

Table 10: Posterior summary measures of the ratio of the vaccine coverage in a given Geography/Insurance group stratum compared to the corresponding Insurance group in North Carolina

		Summary measures			95% HPD interval	
Geo.	Ins.	Mean	Mode	SD	LL	UL
FL	Medicaid	0.996	0.996	0.014	0.968	1.024
FL	Private	0.997	0.997	0.010	0.978	1.016
FL	Uninsured	0.989	0.988	0.040	0.911	1.067
GA	Medicaid	0.983	0.983	0.015	0.953	1.012
GA	Private	0.989	0.989	0.010	0.968	1.008
GA	Uninsured	0.955	0.955	0.040	0.876	1.035
MS	Private	0.977	0.978	0.015	0.948	1.005
MS	Uninsured	0.914	0.914	0.054	0.805	1.019
WI	Medicaid	0.957	0.957	0.016	0.925	0.989
WI	Private	0.971	0.971	0.011	0.948	0.992
WI	Uninsured	0.889	0.890	0.042	0.806	0.972

Question 9

> Make a caterpillar plot of the estimated coverage (per location and insurance status). Include also the observed vaccination proportion in the plot.

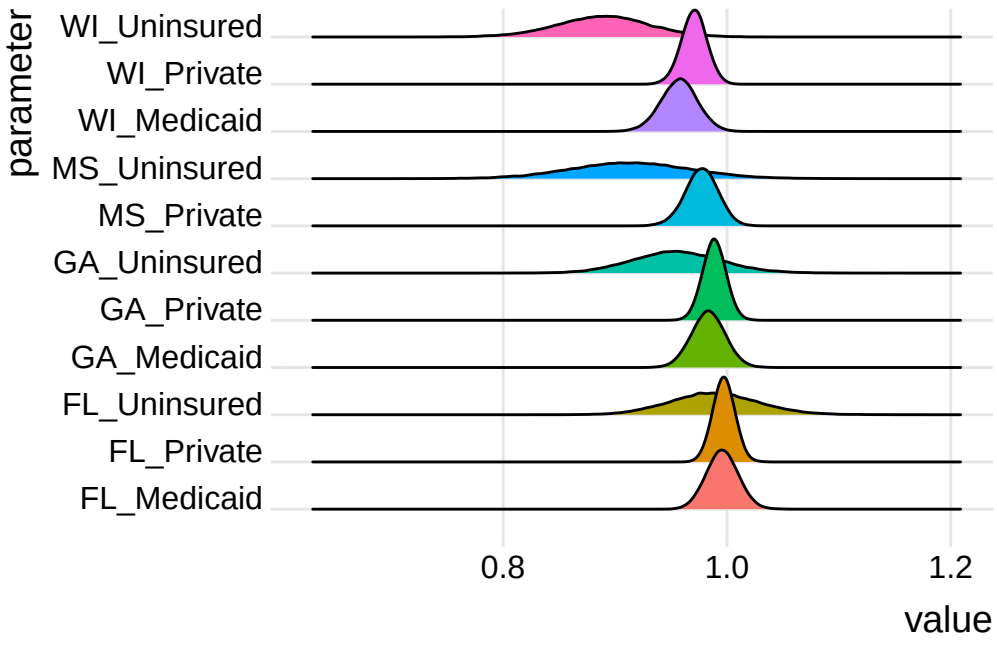
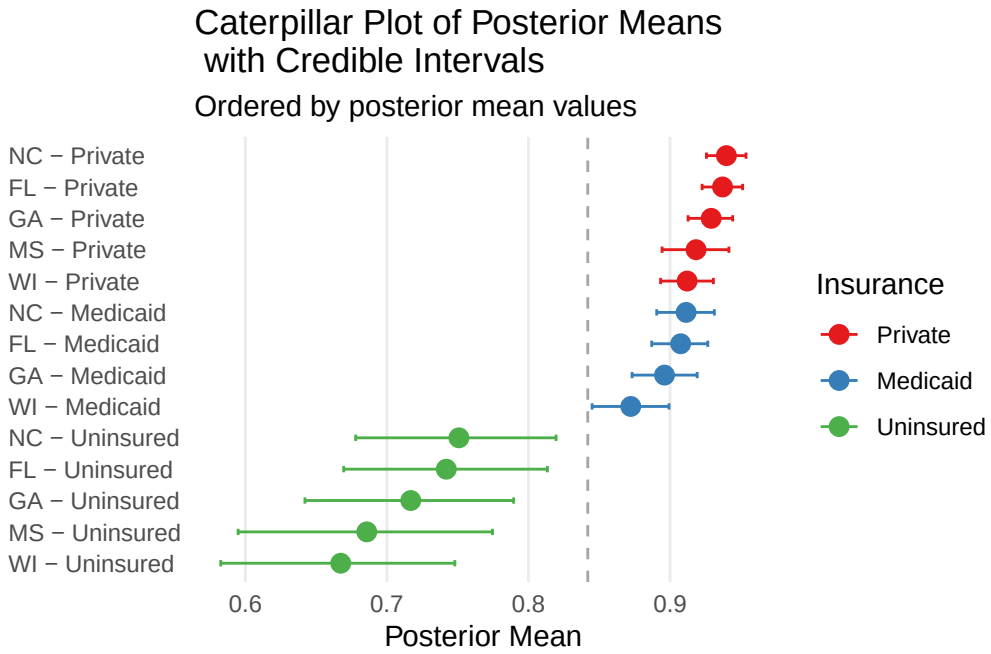


Figure 10: Ridgeline plot showing the posterior distribution of the ratio of the vaccine coverage in a given Geography/Insurance group stratum compared to the corresponding Insurance group in North Carolina



Question 10

> No data is given for children insured by Any medicaid in Mississippi. Predict the number of vaccinated individuals in Mississippi among 519 children with any medicaid

The posterior predictive distribution of the sample outcome in for MS/Medicaid with a sample size of $n = 519$ can be obtained using the posterior distribution of the logodds obtained from the additive logistic regression model built in Question 10.

Posterior sample of the vaccine coverage

m : Contrast row vector for the Mississippi and Medicaid insurance group

$\theta^{(s)}$: Row vector of posterior samples of model parameters (rows: MCMC iteration, columns: sampled parameter values) $s = 1, 2, ..., S$

$\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$

1. Calculation of logodds $\eta^{(s)}$

$$\eta^{(s)} = m\theta^{(s)\top}$$

2. Inverse logit transformation

$$\pi^{(s)} = \text{logit}^{-1}\left(\eta^{(s)}\right)$$

3. Binomial sampling

$$y^{(s)} = \mathcal{B}in\left(n = 519, \pi^{(s)}\right)$$

The posterior predictive distribution of y is approximated by $\left\{y^{(s)}\right\}_{s=1}^{S}$.

Table 11: Summary measures and HPD interval for the posterior predictive distribution for vaccinated people in a sample of size 519 in the MS/Mississippi stratum.

Summary measures			95% HPD interval	
Mean	Median	SD	LL	UL
457.117	458	12.87	429	479